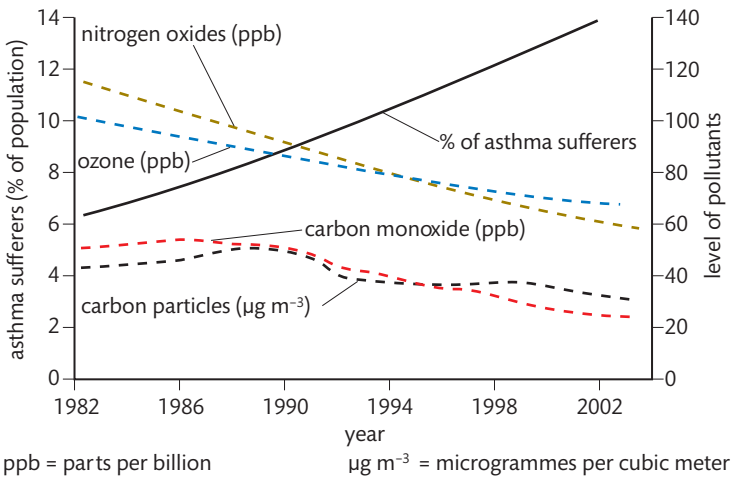
Mathematics, and information and communication technology skills

An important part of being a good scientist and of being a citizen in today's information-rich world is to be number-savvy and tech-savvy. This chapter is divided into two parts: Part 1 Mathematics and statistical analysis, and Part 2 Information and communication technology (ICT) in biology.

In Part 1, we will explore the kinds of mathematical skills needed by a student of biology in order to understand some basic operations and ways of statistically analysing scientific data. Hopefully this section will make you to feel more comfortable with data and give you strategies for understanding graphs and statistical tests that will improve both your internal assessment (IA) work and your exam results.

In Part 2, we will look at how computers, tablets, data-logging devices, and software programs can help us work with numbers and statistics, notably for lab reports.



ppb = parts per billion

μg m^{-3} = microgrammes per cubic meter

Figure 1 Large quantities of data give us superpowers: they allow us to see things other people cannot see. Being able to collect and process data are important skills but also students need to know how to interpret data, including reading graphs, grasping statistics and understanding units and their uncertainties. This graph contains an impressive amount of information in just a few square centimetres – there are 20 years of measurements of five different things. The graph aims to answer the question of whether or not there is a link between asthma and air pollution. Try out your data analysis skills and yours TOK critical thinking skills on this graph.

1 Mathematics and statistical analysis

In the first part of this chapter, you will learn how scientists analyse the evidence they collect when they perform experiments. You will be designing your own experiments, so this information will be very useful to you. You will be learning about:

- means
- error bars
- *t*-tests
- standard deviation
- significant difference
- causation and correlation.

Have your calculator with you to practise calculations for standard deviation and *t*-tests, so that you can use these methods of analysing data when you do your own experiments.

Mean

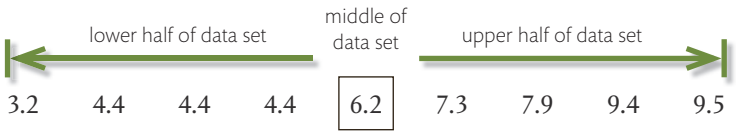
The mean is an average of data points. For example, suppose the height of bean plants grown in sunlight is measured in centimetres (cm) 10 days after planting. The heights for nine of the plants are shown below in cm. The sum of the heights is 56.7 cm. Divide 56.7 by 9 to find the mean (average). The mean is 6.3 cm. The mean shows the central tendency of the data.

3.2 9.5 4.4 6.2 7.9 4.4 9.4 7.3 4.4

In the ICT section of this chapter, you can learn how to use a spreadsheet to calculate the mean and many other values for your data.

Median

Simply put, the median is the number in the middle. It is the number that separates the higher half of the data from the lower half of the data. To find it, the data must first be put in order from the lowest to the highest value. The nine heights of plants have been put in order below.



Because there are nine values, the fifth value separates the lower four from the top four. In cases when there is an even number of data points, take the mean (average) of the two numbers in the middle. For example, if a tenth plant was added to the sample and it was the tallest at 9.7 cm, then the two values in the centre would be 6.2 and 7.3. The mean of those two is 6.75, so the median would be 6.75 cm. But in the example above with nine plants, the median is 6.2 cm.

Mode

The mode is the most frequently occurring measurement. In this case, 4.4 is repeated three times, and no other value is repeated, so the mode is 4.4.

Range

The range is the measure of the spread of data. It is the difference between the largest and the smallest observed values. In our example, the range is 9.5 – 3.2 = 6.3. The range for this data set is 6.3 cm. If one data point was unusually large or unusually small, this very large or small data point would have a big effect on the range. Such very large or very small data points are called outliers. In our sample there is no outlier. If one of the plants died early and had a height of only 0.5 cm, it would be considered to be an outlier. In a lab report, it is acceptable to exclude an outlier from data processing, but it is important to declare it and explain why it was excluded.

Error bars

Error bars are a graphical representation of the variability of data. Error bars can be used to show either the range of data or the standard deviation (SD) on a graph. Standard deviation is explored further on the next page. Notice the error bars representing standard deviation on the bar chart in Figure 2 and the graph in Figure 3.

The value of the standard deviation above the mean is shown extending above the top of each bar of the chart, and the same standard deviation below the mean is shown extending below the top of each bar of the chart. As each bar represents the mean of the data for a particular tree species; the standard deviation for each type of tree will be different, but the value extending above and below a particular bar will be the same. The same is true for the line graph. As each point on the graph represents the mean data for each day, the bars extending above and below the data point are the standard deviations above and below the mean.

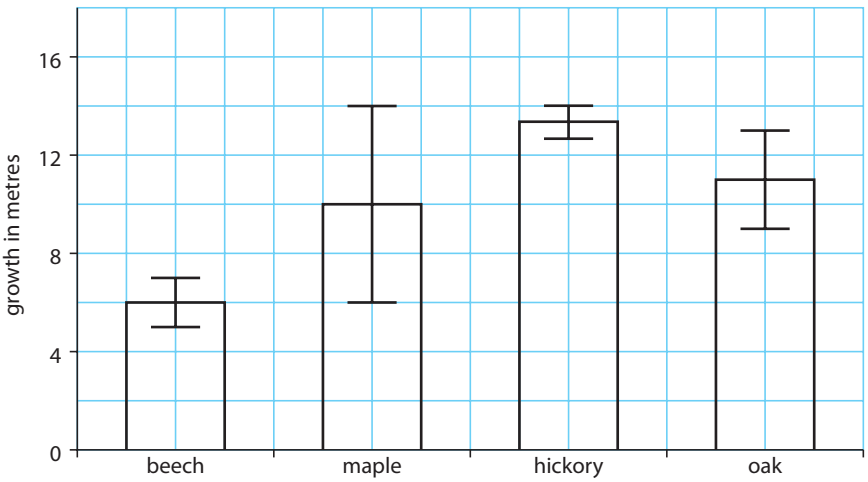


Figure 2 Rate of tree growth on an oak–hickory dune in 2004–05. Values are represented as mean ±1 SD from 25 trees per species.

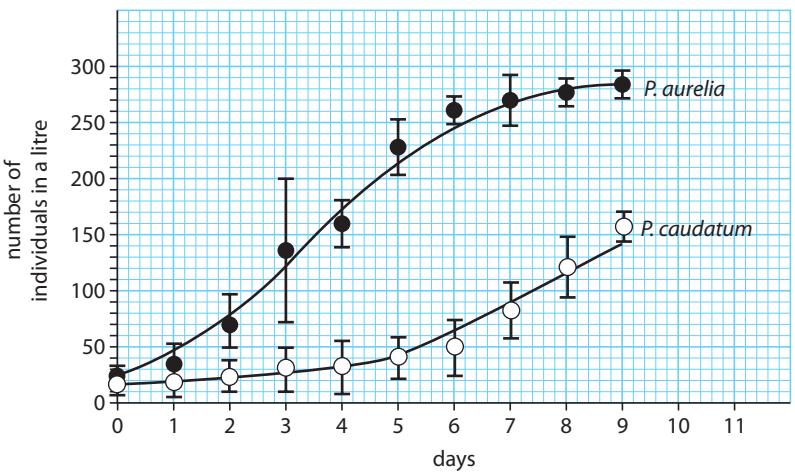


Figure 3 Mean population density ±1 SD of two species of *Paramecium* grown in solution.

Standard deviation

We use standard deviation to summarize the spread of values around the mean, and to compare the means and spread of data between two or more samples. Think of the standard deviation as a way of showing how close your values are to the mean.

In a normal distribution, about 68% of all values lie within ±1 standard deviation (SD) of the mean. This rises to about 95% for ±2 standard deviations from the mean.

To help understand this difficult concept, let's look again at the bean plants. Some bean plants were grown in sunlight, and some were grown in shade. Regarding the bean plants grown in sunlight: suppose our sample is 100 bean plants. Of those 100 plants, you might guess that a few will be very short (maybe the soil they are in is slightly

sandier). A few may be much taller than the rest (possibly the soil they are in holds more water). However, all we can measure is the height of all the bean plants growing in the sunlight. If we then plot a graph of the heights, the graph is likely to be similar to a bell curve (see Figure 4). In this graph, the number of bean plants is plotted on the *y*-axis and the heights, ranging from short to medium to tall, are plotted on the *x*-axis.

Many data sets do not have a distribution that is as perfect as the middle part of Figure 4. Sometimes, the bell-shape is very flat. This indicates that the data are spread out widely from the mean. In some cases, the bell-shape is very tall and narrow. This shows that the data are very close to the mean and not spread out.

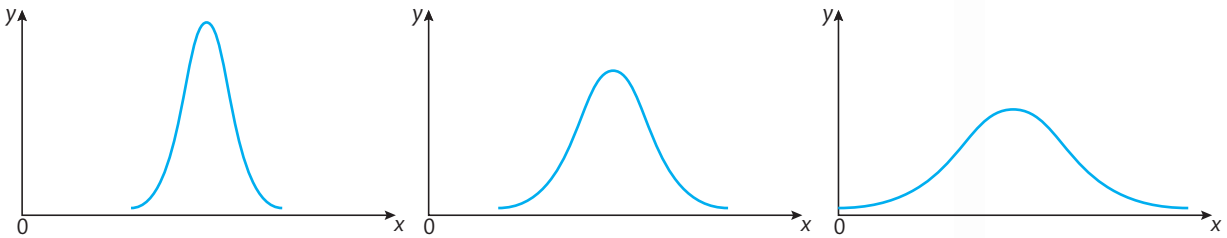


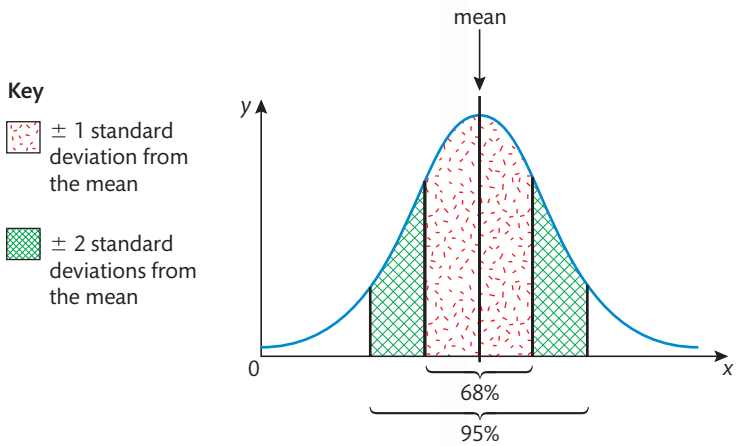
Figure 4 Three different normal distribution curves. The first shows very little spread from the mean, the second shows a moderate amount of spread from the mean, and the third shows a wide distribution of data points from the mean.

The standard deviation shows us how tightly the data points are clustered around the mean. When the data points are clustered together, the standard deviation is small; when they are spread apart, the standard deviation is large. Calculating the standard deviation of a data set is easily done on a calculator with mathematical functions.

For example, if all the students in a year group got 5s and 6s as their marks in a test, the standard deviation would be low, whereas if the results ranged from 2s to 7s, the standard deviation would be higher.

Look at Figure 5. This graph of normal distribution may help you understand what standard deviation really means. The dotted area represents one standard deviation (1 SD) in either direction from the mean. About 68% of the data in this graph are located in the dotted area. Thus we say that, for normally distributed data, 68% of all the values lie within ±1 SD from the mean. Two standard deviations from the mean (the dotted and the cross-hatched areas combined) contain about 95% of the data. If this bell curve was flatter, the standard deviation would have to be larger to account for the 68% or 95% of the data set. Now you can see why standard deviation tells you how widespread your data points are from the mean of the data set.

Figure 5 This graph shows a normal distribution.



How is knowing this useful? For one thing, it tells you how many extreme values are in the data. If there are many extremes, the standard deviation will be large; with few extremes the standard deviation will be small. When processing your data for lab reports, calculating the standard deviation can help you to analyse the data.

Standard deviation is used as an indication of how variable data are. It helps to answer the question 'How far do my data stray from the average?' Data distributed on a normal distribution curve in such a way that most of the data points are near the mean have a low standard deviation (minimal variation), whereas data spread out far from the mean have a high standard deviation (wide variation). Standard deviation can sometimes be used in data processing to help you to decide whether your data are following a clear pattern or whether something is generating unexpected variations.

Comparing the means and spread of data between two or more samples

Remember that in statistics we make inferences about a whole population based on just a sample of the population. Let's continue using our example of bean plants growing in the sunlight and shade to determine how standard deviation is useful for comparing the means and the spread of data between two samples. Table 1 shows the raw data sets at the end of the experiment looking at bean plants grown in sunlight and in shade.

Table 1 Data from a bean plant experiment

Height of 10 bean plants grown in sunlight, in centimetres ±1 cm	Height of 10 bean plants grown in shade, in centimetres ±1 cm
125	131
121	60
154	160
99	212
124	117
143	65
157	155
129	160
140	145
118	95
Total 1310	Total 1300

Bean plants being grown for an experiment.

First, we determine the mean for each sample. As each sample contains 10 plants, we can divide the sum of all the heights by 10 in each case. The resulting means are 131 and 130 cm, respectively.

Of course, that is not the end of the analysis. Can you see that there are large differences between the two sets of data? The heights of the bean plants grown in the shade are much more variable than those of the bean plants grown in the sunlight. The means of each data set are very similar, but the variation is not the same. This suggests that other factors may be influencing growth, in addition to sunlight and shade.

How can we mathematically quantify the variation that we have observed? Fortunately, your calculator should have a function that will do this for you. All you have to do is input the raw data. As practice, find the standard deviation of each raw data set above before you read on.

The standard deviation of the bean plants grown in sunlight is 17.68 cm, while the standard deviation of the bean plants grown in shade is 47.02 cm. Looking at the means alone, it appears that there is little difference between the two sets of bean plants. However, the high standard deviation of the bean plants grown in the shade indicates a very wide spread of data around the mean. The wide variation in this data set makes us question the experimental design. What is causing this wide variation in data? Is it possible that the plants in the shade are also growing

in several different types of soil? This is why it is important to calculate the standard deviation, in addition to the mean, of a data set. If we looked at only the means, we would not recognize the variability of data seen in the shade-grown bean plants.

Significant difference between two data sets using a *t*-test

In order to determine whether or not the difference between two sets of data is a significant difference, *t*-tests are commonly used. The Student's *t*-test (named after a scientist publishing his work under the pseudonym 'Student') compares two sets of data, for example the heights of the bean plants grown in sunlight and the heights of bean plants grown in shade. Look at the top of the table of *t*-values (Table 2) and you will see the probability (*p*) that chance alone could make a difference. If $p = 0.50$, it means the difference could be the result of chance alone 50% of the time.

Statistical significance refers to how probable it is that a relationship is caused by pure chance. If a relationship is statistically significant, it means that there is very little chance that the relationship is caused by chance. We can also use this idea to see whether the differences between two populations are random or not.

For example, a value of $p = 0.50$ (or 50%) is not a significant difference in statistics. It means that there is a 50% probability that the differences are caused by chance alone. However, if you reach $p = 0.05$, the probability that the difference is caused by chance alone is only 5%. This means that there is a 95% likelihood that the difference has been caused by something besides chance. A 95% probability is statistically significant in statistics. Statisticians are rarely completely certain about their findings, but they like to be at least 95% certain of their findings before drawing conclusions.

The formula when comparing two populations that are assumed to have equal variance is as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Note: you will *not* be asked this formula on exams – it is presented here only as something that might be useful for processing the data collected in your laboratory investigations.

$\bar{X}_1$ = the mean of population 1
$\bar{X}_2$ = the mean of population 2
N = sample size of the population
s = standard deviation

If you plug in the values from the above example with bean plants, you should get $t = 0.06$. You can use a table of critical *t*-values (Table 2) to find out what this number means. To do this, look in the left-hand column of Table 2, headed 'Degrees of freedom', then look across to the given *t*-values. For a two-sample *t*-test like the one we are doing, the degrees of freedom (d.f.) are the sum of the sample sizes of the two groups minus two: 10 + 10 - 2 = 18.

If d.f. = 18, we need to look at the row on the table of *t*-values that corresponds to 18. We see that our calculated value of *t* (0.06) is less than 0.69 on the table, indicating that the probability that the differences between the two populations of plants are due to chance alone is greater than 50%. In other words, we can safely declare that there is no statistically significant difference in the data collected from the bean plants in the sunlight and those from the shade. The differences are most likely due to chance. In order to be able to declare that our two populations showed a level of 95% significance in their differences, we would need a *t* value of 2.10 or more (see d.f. = 18 and $p = 0.05$ (5%) in Table 2). Interpretations of such data processing can be a crucial addition to an effective conclusion on a lab report.

When something is considered to be statistically significant, it means that there is a strong probability that it is *not* caused by chance alone. When something could be caused by chance, we say in statistics that it is not statistically significant.

Table 2 ***t*-values**

		Probability (*p*) that chance alone could produce the difference					
		0.50 (50%)	0.20 (20%)	0.10 (10%)	0.05 (5%)	0.01 (1%)	0.001 (0.1%)
Degrees of freedom	1	1.00	3.08	6.31	12.71	63.66	636.62
	2	0.82	1.89	2.92	4.30	9.93	31.60
	3	0.77	1.64	2.35	3.18	5.84	12.92
	4	0.74	1.53	2.13	2.78	4.60	8.61
	5	0.73	1.48	2.02	2.57	4.03	6.87
	6	0.72	1.44	1.94	2.45	3.71	5.96
	7	0.71	1.42	1.90	2.37	3.50	5.41
	8	0.71	1.40	1.86	2.31	3.37	5.04
	9	0.70	1.38	1.83	2.26	3.25	4.78
	10	0.70	1.37	1.81	2.23	3.17	4.59
Degrees of freedom	11	0.70	1.36	1.80	2.20	3.11	4.44
	12	0.70	1.36	1.78	2.18	3.06	4.32
	13	0.69	1.35	1.77	2.16	3.01	4.22
	14	0.69	1.35	1.76	2.15	2.98	4.14
	15	0.69	1.34	1.75	2.13	2.95	4.07
	16	0.69	1.34	1.75	2.12	2.92	4.02
	17	0.69	1.33	1.74	2.11	2.90	3.97
	18	0.69	1.33	1.73	2.10	2.88	3.92
	19	0.69	1.33	1.73	2.09	2.86	3.88
	20	0.69	1.33	1.73	2.09	2.85	3.85
	21	0.69	1.32	1.72	2.08	2.83	3.82
	22	0.69	1.32	1.72	2.07	2.82	3.79
	24	0.69	1.32	1.71	2.06	2.80	3.75
	26	0.68	1.32	1.71	2.06	2.78	3.71
	28	0.68	1.31	1.70	2.05	2.76	3.67
	30	0.68	1.31	1.70	2.04	2.75	3.65
	35	0.68	1.31	1.69	2.03	2.72	3.59
	40	0.68	1.30	1.68	2.02	2.70	3.55
	45	0.68	1.30	1.68	2.01	2.70	3.52
	50	0.68	1.30	1.68	2.01	2.68	3.50
	60	0.68	1.30	1.67	2.00	2.66	3.46
	70	0.68	1.29	1.67	1.99	2.65	3.44
	80	0.68	1.29	1.66	1.99	2.64	3.42
	90	0.68	1.29	1.66	1.99	2.63	3.40
	100	0.68	1.29	1.66	1.99	2.63	3.39

Worked example

Two groups of barnacles living on a rocky shore were compared. The width of their shells was measured to see whether there was a significant size difference depending on how close they lived to the water. One group lived between 0 and 10 m above the water level. A second group lived between 10 and 20 m above the water level.

The width of the shells was measured in millimetres (mm). Fifteen shells were measured from each group. The mean size of the group living closer to the water indicated that barnacles living closer to the water had larger shells. If the value of t is 2.25, is that a significant difference?

Solution

For one of the steps of the Student's t-test, we need to determine the degrees of freedom. In an example like this one, where the two sample sizes are equal and we can assume the variance in the two samples is the same, the degree of freedom is $2n - 2$. The letter n represents the sample size (the number of measurements made), and in this case $n = 15$. The degrees of freedom in this example is 28 because $(2 \times 15) - 2 = 28$. Looking along the row of Table 2 that shows the degrees of freedom of 28, we see that 2.25 is just above 2.05.

Referring to the top of this column in the table, $p = 0.05$: so the probability that chance alone could produce that result is only 5%.

The confidence level is 95%. We are 95% confident that the difference between the barnacles is statistically significant. In other words, the differences in mean size is very unlikely to be a product of pure chance.

Note: when calculating the t-test value using a spreadsheet program such as Microsoft Excel, be aware that the value obtained is the % chance rather than the value for t. As a result, you do not need to look up the critical values in the table.

Correlation does not mean causation

We make observations all the time about the living world around us. We might notice, for example, that our bean plants wilt when the soil is dry. This is a simple observation. We might carry out an experiment to see whether watering the bean plants prevents wilting. Observing that wilting occurs when the soil is dry is a simple correlation, but the experiment provides us with evidence that the lack of water is the cause of the wilting. Experiments provide a test that shows cause. Observations without an experiment can only show a correlation. Also, in order for these to be evidence of causality, there must be a mechanism to explain why one phenomenon might cause the other. Knowing the properties of osmosis and turgidity in plant cells would explain the causality associated with the correlation, thus giving it great scientific plausability.

A cormorant.

Cormorants

When using a mathematical correlation test, the value of the correlation coefficient, r, is a measure of the degree of linear relationship or linear dependence between two variables. This can also be called the Pearson correlation coefficient. The value of r can vary from +1 (completely positive correlation) to 0 (no correlation) to –1 (completely negative correlation). For example, we can measure the size of breeding cormorant birds to see whether there is a correlation between the sizes of males and females that breed together.

Table 3 Cormorant size data

Pair number	Size of female cormorants, cm	Size of male cormorants, cm
1	43.4	41.9
2	47.0	44.2
3	50.0	43.9
4	41.1	42.7
5	54.1	49.5
6	49.8	46.5
$r = 0.88$		

Correlation does not necessarily mean causality. Just because two things show a relationship and have a strong *r*-value, does not mean one causes the other.

The *r*-value of 0.88 shows a positive correlation between the sizes of the two sexes: large females mate with large males. However, correlation is not cause. To find the cause of this observed correlation requires experimental evidence. There may be a high correlation, but only carefully designed experiments can separate causation from correlation. Causality requires that the mechanism of exactly how X causes Y needs to be demonstrated. For example, the mathematics here does not explain whether it is the males choosing the females or the females choosing the males. Correlation says nothing about the direction of the influence.

Graphs

Scientists use graphs extensively because they are useful tools for presenting data and seeing relationships that might otherwise remain hidden. Graphs are instrumental in analysing data, and if you know how to make accurate and appropriate graphs your conclusion and evaluation will be greatly enhanced.

The most common forms of graphs you are expected to be able to use are:

- bar charts
- histograms
- line graphs
- scatter plots.

Occasionally, you may also need to use pie charts or box and whisker diagrams, but here we will focus on the four listed above.

Bar charts

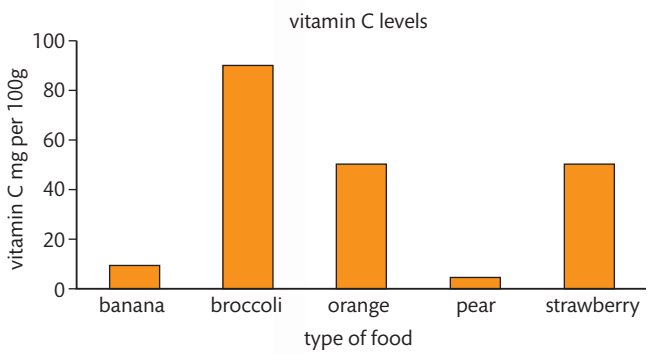


Figure 6 A bar chart showing vitamin C levels in different types of food.

Bar charts use rectangles to show the amount of data in a certain number of categories. The height of each rectangle corresponds to a quantitative value. The *y*-axis is quantitative, but the *x*-axis shows categories rather than incremental numerical values. The order of these categories could be changed and it would not make a difference. Empty spaces separate the rectangles along the *x*-axis. For example, Figure 6 is a bar chart showing the amount of vitamin C in various foods.

In a graph of this type, it is okay to rearrange the bars anyway you want. In Figure 6, the data are presented alphabetically, but there is no reason why you couldn't order the bars from the greatest to the smallest numerical values.

Histograms

Histograms have some similarities with bar charts, except that the *x*-axis has a quantitative scale marking off intervals of continuous data. In addition, the widths of the rectangles that make up the histogram represent specific incremental quantitative values. The histogram in Figure 7 shows the amount of time that 42 individuals of a particular species of animal spent drinking at a river.

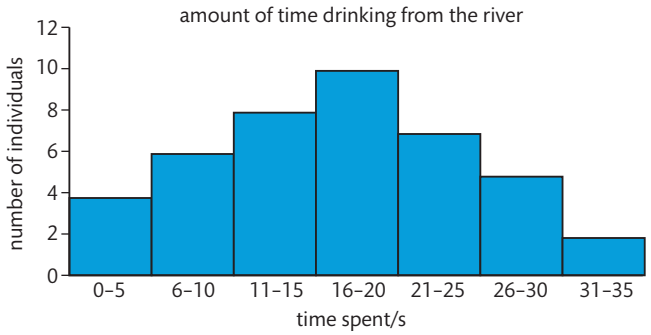


Figure 7 A histogram of the time spent by individuals of a species drinking from a river. Notice the lack of space between the categories, and the fact that the categories on the *x*-axis represent continuous incremental numerical values.

Histograms have no spaces between the rectangles because the data are continuous. This was not the case in the bar chart that we looked at in Figure 6. In Figure 7, we cannot rearrange the rectangles of the histogram so that the highest values are on the left and the lowest values are on the right, as we could have done for the bar chart. Histograms must follow the scale shown on the *x*-axis. If an animal drank for 24 seconds, the data must go in the range 21–25. These ranges can also be called bins, and you can think of a histogram as a series of bins that you fill up with the appropriate data as the data are sorted.

Line graphs

A line graph plots single points over regular increments such that each *x*-value has only one corresponding *y*-value. The dots are then joined with straight lines. The example in Figure 8 shows a newborn baby's body mass between the time of its birth and the age of 18 months.

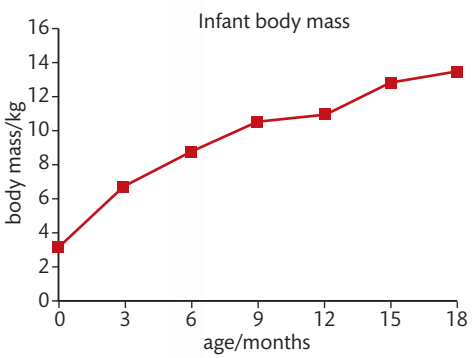


Figure 8 A line graph of infant body mass. Notice that the data points are connected by straight lines rather than using a trend line or line of best fit.

In line graphs, the *x*-axis is usually the independent variable, in which case the *y*-axis is the dependent variable. There is clearly a correlation in this graph: as age increases, body mass increases. There is a positive correlation. But remember, that does not mean there is causality. Ageing is not the mechanism that causes an increase in the child's body mass; on the contrary, good nutrition, genes, and growth hormones are more likely candidates for causing the increase. Line graphs can sometimes show discrepancies in the data. For example, a doctor might wonder why a child did not grow as fast between the ages of 9 and 12 months compared with the rest of the graph. Perhaps the child did not have access to proper nutrition during that interval.

Scatter plots

A scatter plot is used when two variables are involved, and they are plotted as *y* against *x* using Cartesian coordinates. Such graphs work well for situations where one *x*-value may have multiple *y*-values. As with line graphs, scatter plots are useful for trying to see a correlation. Figure 9 shows a scatter plot for the numbers of pairs of grey partridges (a type of bird) plotted against the number of sightings of birds of prey per square kilometre (km^{-2}).

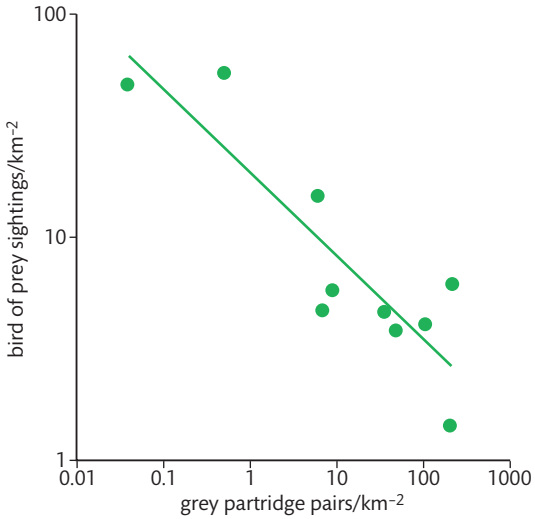


Figure 9 A scatter plot of grey partridge pairs against bird of prey sightings, with logarithmic scales on the *x*- and *y*-axes. M. Watson et al. 2007

Notice how the dots are a bit irregular: this scattering is where this type of graph gets its name from. Notice also how the data points are *not* connected by a line. Rather, a line of best fit or a trend line has been placed over the graph showing an overall trend in the data. Such lines or curves do not need to pass through each data point, as we saw in the line graph. The trend line in Figure 9 shows that there is a negative correlation. A negative correlation means that as one variable increases the other decreases.

Do you notice anything peculiar about the axes? They are shown using a logarithmic scale, which means that each increment is 10 times the size of the one before. This is relatively exceptional: most scatter plots have standard incremental scales on the *x*- and *y*-axes, the way the line graph does in Figure 8. Logarithmic scales are useful when you are trying to show distributions of data points that would not show up if they were put on a normally incremental scale. In this case, it is likely that the authors of the report in which this graph appeared wanted to show the correlation between the sightings of birds of prey and the number of couples of partridges. As we know that there is a logical mechanism for causality (birds of prey kill and eat partridges), it is not impossible to suspect that there is a causal relationship here. But this graph alone cannot prove that birds of prey cause the reduction in numbers of partridges.

Regression models and coefficient of correlation

When scientists measure something, often they are looking to see whether they can demonstrate that the phenomenon is following a law of nature. Sometimes laws of nature follow patterns that can be expressed in mathematical equations. For example, when measuring the light that a leaf might use for photosynthesis, a scientist knows that the intensity of the light varies according to an equation relating intensity with the

distance to the light. You know from personal experience that holding a torch close to your eyes can be blinding, whereas seeing the same torch from far away does not hurt your eyes. In Figure 10, the graph on the left illustrates the 'pure' mathematical law about light intensity and distance from the light source. On the right is the same graph superimposed with measurements taken in a lab. Because of any number of things, including limitations in the equipment and human error, the lab measurements do not fit the mathematical model perfectly.

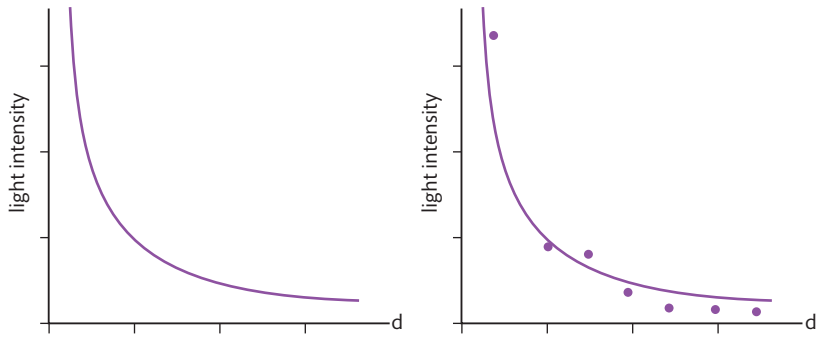
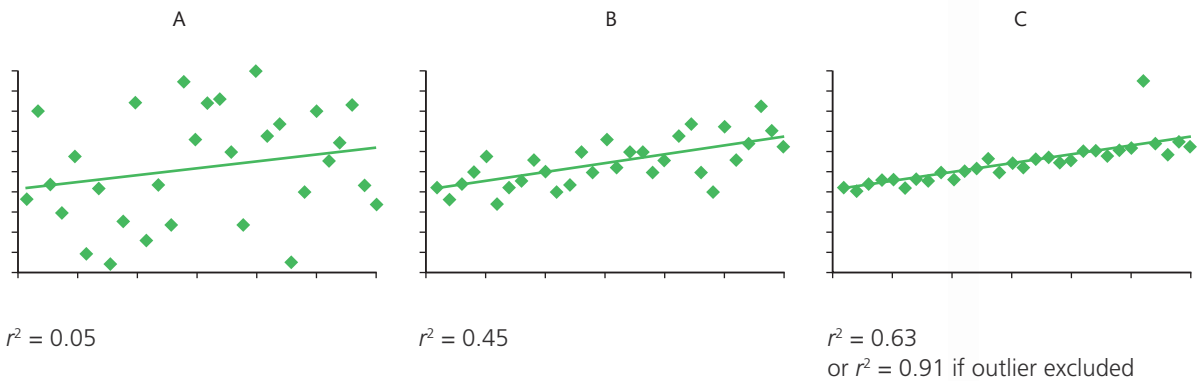


Figure 10 A model of what the data should show (on the left) and the actual collected data (dots on the right), of which only one point is actually where it was expected to be. The d on the *x*-axis of the graphs represents the distance from the light source.

Now imagine the opposite. A scientist takes some measurements and wonders if there is some kind of mathematical equation that could act as a model of her data. She makes a scatter plot and then sees if there is a trend line that fits her data reasonably well. She might start with a straight line because that is the simplest relationship between two variables. This is called a simple linear regression model. But if that does not fit her data well, she could try other regression models that are not straight lines. Fortunately, statistical functions in her calculator or spreadsheet program on her computer can do this for her in an automated fashion.

How can we know if the trend line's regression model is the best one for the data we collected? The squared correlation coefficient, r^2, also called the coefficient of determination, is used to see how well a regression model matches the data collected. A value of $r^2 = 0$ means the regression model does not fit the data at all, whereas a value of $r^2 = 1$ means a perfect fit. Note that r^2 cannot be a negative number. Here are some examples showing the r^2 values calculated by Microsoft Excel for three data sets and their trend lines.

Figure 11 Three examples of data that have been modelled with a linear regression. The r^2-value is then calculated to see how closely the linear regression model matches the data.



Notice what happens to the r^2-value as the variability of the data points is reduced from A to B to C. This reveals that the regression model shown by the trend line matches the observed data better and better. Graph A's regression line suggests that there is very

- Trend lines are useful for seeing whether there is an overall pattern or tendency in the data points.
- The r^2-value, the coefficient of determination, is useful for seeing if the trend line matches the data points closely or not. It indicates how good the model is. The closer it is to 1, the better the model. Values close to 1 reveal that there is a strong correlation between the *x*- and *y*-values.
- If the regression model fits the data well, it can be used to predict values that were not measured.

little evidence of an agreement between the regression model and the data, whereas B and C show a stronger fit. Notice what happens in graph C: there is clearly an outlier at the top right. Fortunately, the investigator identified it as being a result of an error during the lab. It can safely be ignored, and therefore the value of 0.91 can be used for analysis purposes. Students are encouraged to used trend lines and r^2-values in their data processing, in order to analyse the data they have collected better.

In addition to simply seeing whether the data points follow a predictable pattern, a regression model can be used to predict values that were not measured. Knowing the equation of the line or the curve allows a researcher to plug in hypothetical values and get a prediction from the model. For example, changes in the human population in the coming decades can be predicted based on a regression model of current trends in the population. When using a regression model for prediction purposes, the r^2-value can help give a sense of how reliable the prediction will be. For example, predicting an outcome using graph A above would be extremely unreliable. However, using C's regression model would be more likely to give reliable results.

Before and after: by how much did this change?

Sometimes we need to analyse how something has changed over time, or we need to see whether there is a difference between what we expected and what we got.

The simplest way to see a difference is to subtract the 'after' value, V_2, from the 'before' value, V_1. However, it is often practical to calculate a percentage change:

$$\text{percentage change} = \left(\frac{V_2 - V_1}{V_1}\right) \times 100$$

Expected versus observed values: first application of the chi-squared test for goodness of fit

As we saw in Figure 10, we do not always get what we expect with our results. The difference between the expected values and the observed values may simply be caused by chance or, on the contrary, may be because an unexpected phenomenon is having an effect on the data. How can we know? One way to answer this question is to carry out a statistical test called the chi-squared (χ^2) test, which calculates how close our observed results are to the expected values. Chi is the Greek letter χ and is pronounced like the word 'sky' without the s at the beginning.

The first way we will use the χ^2 test is to compare our observed results with what we can theoretically calculate the results should be (the 'expected' results). As you saw in Chapter 10, to use this statistical test it is important to note down carefully all the observed results (O) and the expected results (E). In the case of genetics exercises, the expected results would be the proportions of phenotypes as determined by a Punnett grid, such as 25%/50%/25% or 25%/75%, although it is important to use the actual numbers of offspring rather than percentages or ratios. Setting up a table to help keep track of the numbers is helpful.

Table 4 Charting observed and expected results

	Possible outcome 1	Possible outcome 2	Sum
Observed numbers in each category of possible outcomes (O)			
Expected numbers in each category of possible outcomes (E)			
Difference (O – E)			
Difference squared $(O - E)^2$			
$\frac{(O - E)^2}{E}$			$\chi^2 =$

The third and fourth lines of this table are intermediate steps to see the difference between the observed and the expected values as well as their squared values.

The bottom right cell of the table is what we want: it shows the sum of the last row's values and this is the χ^2 value we are interested in. In effect, the contents of this table can be summarized in the generalized formula for calculating χ^2, which is:

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

where χ = Greek letter chi, O = observed values (the results of the experiment), E = expected values (calculated theoretically), Σ = sum of all the calculations for each type of outcome.

Interpreting the χ^2 value calculated

Once we know the χ^2 value, we need to know what it means. For this there are some concepts that need to be clarified. First of all, there is the concept of the null hypothesis (H_0). The H_0 in an experiment of this type is what we would expect: this is usually determined by mathematical calculations. The χ^2 value will help us to determine whether the null hypothesis can be rejected. Accepting the null hypothesis is a way of saying 'Yes, there is a high probability that any deviation from the expected values can be attributed to chance'.

Another two important concepts to understand are the idea of degrees of freedom (d.f.) and how the idea of probability (*p*) is used. When using the χ^2 test to determine whether there is a difference between the expected and the observed values, the degrees of freedom is determined by taking the number of categories into which the data fall and subtracting 1 from that number. In Table 4, there are two categories into which the data fall (possible outcomes 1 and 2, so there is 2 – 1 = 1 degree of freedom). This number allows us to know where to look in the table of critical values for χ^2 (Table 5). Notice in Table 5 that, in addition to the degrees of freedom, there are probability values for *p*. It is a convention in biology to look for probabilities of 5%, or 0.05.

Do not confuse Tables 2 and 5. Although they both refer to probability and hypothesis testing, the former is for the *t*-test and this one is for the chi-squared test.

Table 5 Critical values for χ^2

		Probability values (p)				
		0.1	0.05	0.025	0.01	0.005
Degrees of freedom (d.f.)	1	2.706	3.841	5.024	6.635	7.879
	2	4.605	5.991	7.378	9.21	10.597
	3	6.251	7.815	9.348	11.345	12.838
	4	7.779	9.488	11.143	13.277	14.86
	5	9.236	11.07	12.833	15.086	16.75
	6	10.645	12.592	14.449	16.812	18.548
	7	12.017	14.067	16.013	18.475	20.278
	8	13.362	15.507	17.535	20.09	21.955
	9	14.684	16.919	19.023	21.666	23.589
	10	15.987	18.307	20.483	23.209	25.188

Look at Table 5 and find the critical value that is of interest to us: it is the one that lines up with a probability value of 0.05 and a degree of freedom of 1. You should get 3.841. This means that any value we calculate for χ^2 that is greater than 3.841 tells us to reject the null hypothesis.

Here is a summary of the steps.

1. Determine the expected values (although we sometimes like to use percentages or proportions in science, the χ^2 test requires numbers here: do not use percentages or ratios).
2. Note down the observed values and decide what the null hypothesis will be.
3. Calculate the value for χ^2 by determining the differences between the values $(O - E)$, then square them, $(O - E)^2$, and finally add them all up.
4. Determine the degrees of freedom (d.f.) by taking the total number of classes into which the data fall and subtracting 1.
5. Look at the table of critical values of χ^2 and use the d.f. and *p*-value (conventionally we use 0.05 for *p*) to determine which critical value ($\chi^2_{critical}$) to compare the calculated value of χ^2 ($\chi^2_{calculated}$) to.
6. Compare $\chi^2_{critical}$ to $\chi^2_{calculated}$ and decide if the null hypothesis can be rejected using these rules:

$\chi^2_{calculated} < \chi^2_{critical} < \chi^2_{calculated}$

do not reject null hypothesis,

any deviations from the expected values are probably the result of chance alone

reject null hypothesis,

deviations from the expected values are *not* the result of chance alone

If the calculated value for χ^2 is less than the critical value, the null hypothesis cannot be rejected, whereas if the calculated value for χ^2 is greater than the critical value, the null hypothesis can be rejected.

Independent or correlated: second application of the chi-squared test as a test for independence

As seen in Chapter 4, sometimes we need to know whether it is likely that two phenomena are independent from each other or associated with each other. This next application of the χ^2 test will also compare expected and observed values, but this time the expected frequencies are not given in advance. The use of a contingency table like the one below is necessary to determine them. Table 6 shows the data relevant to the quadrat experiment described in Chapter 4, in which students wanted to see whether the distribution of ferns was random or whether they were found more commonly in sunny or shady areas.

Table 6 Quadrat data

Observed:		Area sampled		
		Sunlight	Shade	
Presence of ferns	Present	7	14	21
	Absent	13	6	19
		20	20	40

The cells in pink show the two columns and two rows of observed data; the yellow cells show the marginal totals for the two rows; and the blue cells show the marginal totals for the two columns. The number 40 represents the whole sample size of 40 quadrats (20 from the sunlit areas, and 20 from the shaded areas).

Determining the expected values

Unlike the previous use of the χ^2 test, we have no mathematical model to predict the theoretical 'expected' values. For that, we construct a new table by removing the observed values.

Table 7 Determining the expected values, step 1

		Area sampled		
		Sunlight	Shade	
Presence of ferns	Present			21
	Absent			19
		20	20	40

Now, to fill in the table with expected values, we multiply the marginal total of each row by the marginal total of each column.

Table 8 Determining the expected values, step 2

Expected:		Area sampled		
		Sunlight	Shade	
Presence of ferns	Present	(20 × 21) ÷ 40 = 10.5	(20 × 21) ÷ 40 = 10.5	21
	Absent	(20 × 19) ÷ 40 = 9.5	(20 × 19) ÷ 40 = 9.5	19
		20	20	40

There are some conditions that need to be met when using the χ^2 test.

- **This kind of statistical test works with data that you can put into categories and you want to find out whether the frequency that the results fall into a particular category is the result of chance alone.**
- **Make sure the categories into which the data can fall are exhaustive and mutually exclusive, such as yes/no, or red flower/white flower/pink flower. As with flipping a coin, heads/tails, all the data collected must fall into one or the other of the categories.**
- **Make sure the data sample is sufficiently large: with fewer than five data points in any one category, the result will not be very reliable.**

The null hypothesis is usually the opposite of the investigator's hypothesis. For example, if a doctor wanted to study the effects of a drug on her patients, she might have the hypothesis 'This drug has a positive influence on my patients' health. Compared with the control group not taking the drug, the experimental group will declare more often that they feel better.' In such a scenario, the null hypothesis would be: 'This drug has no influence on my patients' health. There is no difference between the control group and the experimental group: I can be confident that any observed differences will be due to chance alone.'

This is why researchers are happy and satisfied when they can reject the null hypothesis. They are glad to see that they can rule out the idea that the results are only caused by chance. But be careful: just because the null hypothesis can be rejected, it does not mean that the investigator's hypothesis has been validated.

Degrees of freedom

To determine the degrees of freedom, take the number of rows (r) minus one and multiply that by the number of columns (c) minus one. In this case, there are two columns (sunlight and shade) and two rows (ferns present and ferns absent), so the formula is:

$$\text{d.f.} = (r-1)(c-1)$$

$$\text{d.f.} = (2-1)(2-1) = 1$$

Calculate the chi-squared value

For most tables of contingency, the normal formula for χ^2 can be used:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E}$$

Your calculator or spreadsheet program most likely has formulas to calculate this value quickly, but using tables like Table 4 can allow you to walk through the calculation step by step. You should get 4.91 as the critical value of t.

Test for independence: interpreting the chi-squared value

Now we need to look at the critical values table (Table 5). As we have 1 degree of freedom and we always look at $p = 0.05$, the critical value comes out as 3.841. As our calculated value (4.91) is higher than this critical value, we can safely reject the null hypothesis. What does this mean? It means that if the null hypothesis were true and the distribution of ferns was solely the result of chance, there would be less than a 5% chance of getting the results we observed. The fact that our calculated χ^2 value is high means that the relationship is statistically significant. It also means that the distribution of ferns and the presence of sunlight are not totally independent from each other. We can reject the idea that they are independent.

2 Information and communication technology in biology

In the second part of this chapter, we will look at how digital technology can be applied to biology. Thanks to advances in desktop computers, laptops, tablets, and smartphones, many tools that would have only been available to highly specialized labs a few decades ago are now available to everyone. In the 1980s, for example, three-dimension (3-D) animation was cutting-edge technology requiring a roomful of computer processors. Today, teenagers can sketch objects in 3-D on their smartphones. We will look at the following aspects of Information and communication technology (ICT) that apply to biology:

- models
- simulations
- databases
- questionnaires and surveys
- data-analysis exercises
- fieldwork and data logging
- ICT skills as applied to lab reports.

Models

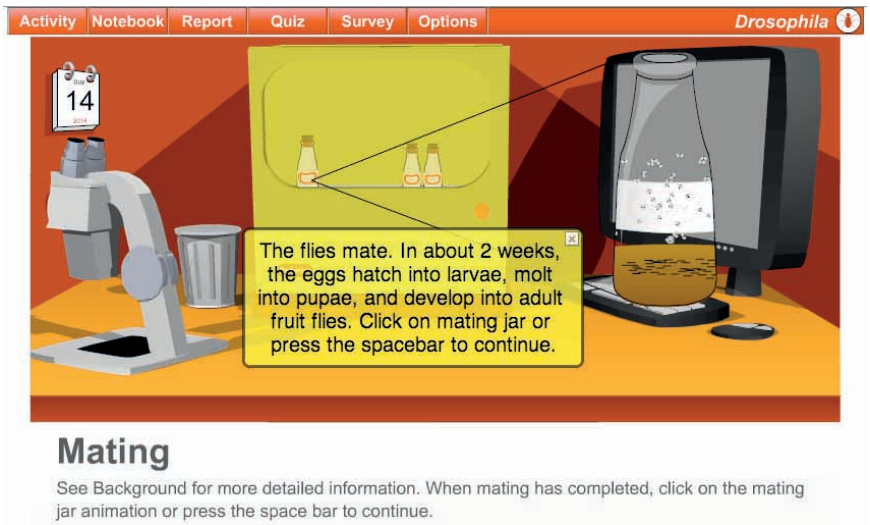
A model is a simplified representation of an object or a phenomenon that can be used to better explain or understand it. Physical models, such as a plastic model of a heart, might help a student to see how the valves work to keep the blood flowing in one direction between the chambers. Computer models, such as a 3-D animation of a beating heart or abstract models showing a flowchart of a process such as DNA replication, can help the learner grasp complex concepts by providing simplified visualizations. Computational models, such as climate models, might be used to help simulate Earth's true climate on a computer.

A computer model can be used to study 3-D objects. Once an object such as this skull is represented as a 3-D model, the data can be shared with other labs and be studied by multiple experts all over the world simultaneously.

Simulations

Models can be applied in simulations in order to represent a process or system. Because variables can be manipulated within them, simulations are often used to predict an outcome or to find out what the optimum parameters are for a system. For example, computer simulations use climate models to predict what will happen to Earth's climate if carbon dioxide levels increase. In the lab, some experiments are too

dangerous, too time-consuming, or too costly to carry out; computer simulations of those experiments can be performed on a computer safely and in a time-saving fashion. Experiments mating fruit flies, for example, would take many weeks to do, or experiments on the effects of introducing predators into an ecosystem would take months or years, and not be very realistic for an IB student to undertake. However, computer simulations can allow a student to collect data and perform experiments virtually on screen. See the hotlinks section at the end of this chapter for examples of online simulations.



A simulation for mating fruit flies to see what kinds of genetic combinations are possible. Notice that this simulation uses models of a microscope, an incubator, glassware, and a lab bench to simulate this experiment with virtual flies that follow a genetic model. sciencecourseware.org

Entering data about blood plasma into a database using a bar code reader.

Databases

Students are often encouraged to compare the values they get in the lab investigations they carry out with values that scientists in other labs have obtained. In other instances, students do not have access to the lab equipment necessary to do certain experiments, such as finding gene sequences or measuring carbon dioxide concentrations over many decades. In either case, databases are available online for a variety of types of data, and students should take advantage of these resources. Hotlinks to some useful databases can be found at the end of the chapter.

Gathering your own statistics

Questionnaires and surveys can sometimes come in handy when students are looking for large quantities of data to analyse. Writing a good survey or questionnaire is an art as well as a science. As with many projects, once you have an idea for your research question, it is best to start with the end in mind and then work backwards.

1. Picture the kinds of graphs that you would want to see on your final data processing that would lead you to an interesting conclusion.
2. Then think of what kinds of data need to be collected in order to produce such graphs. For example, suppose you want to

find out whether the use of flashcards helps students perform better when doing biology multiple-choice questions. You would need to decide whether you want to ask how many flashcards students have made or how much time they spend reviewing them, or both.

3 Because you are trying to show the influence of X on Y, you would probably want to do some kind of scatter plot graph with a trend line to see whether there is a positive or negative correlation. Perhaps you could do some data processing to find out whether the number of cards and/or the amount of time they are used is independent of the students' test scores or not. You could calculate whether there was a statistically significant difference between one group and another in terms of test performance.

4 To see if there is an influence, you would need to obtain the test scores from the participants in your study. This raises some ethical questions because certain students might not want to give you that information. Every time you do a questionnaire, you must tell the students what information is being collected, why it is being collected, and what will be done with the information. They have the right to know, for example, if your data is going to be shared with other people. If your intent is to collect anonymous data, you can reassure your participants that their names will not appear anywhere in the data. Also, you should give the participants the opportunity to leave certain questions blank.

5 Use these helpful hints about setting up questionnaires and surveys.

(a) Even if your questionnaire is anonymous, be sure to collect some demographic information, such as female/male, age, year group, etc. Put such questions at the end of your questionnaire or survey rather than at the beginning. This information might prove useful later because you might see some unexpected trends in the data, such as which age groups use flashcards the most. Some of the best discoveries are the unexpected ones.

(b) So that the data are easier to use in a spreadsheet, use tick boxes or multiple-choice questions whenever possible. Avoid open-ended questions where participants write their own answers. For example, in an open-ended question about gender, some participants may write 'male' and others may write 'M'. A computer would see that as two different answers, even though we know they both mean male.

(c) Be sure that your categories do not overlap. For example, if you are asking students to tick their age group, do not put '13 to 15' as one category and '15 to 17' as another category because students who are 15 will not know which one to tick.

(d) Before you send out your questionnaire to all the participants, try it out on a few classmates and teachers. Often they can spot errors that you did not see or they might have suggestions for clarifying certain points.

Although it is possible to print out and photocopy sheets to collect data, it is very time-consuming to type in all the answers into a spreadsheet when you want to do your data processing. It is much quicker to set up an online questionnaire so that as soon as participants click 'submit' the answers are added to a spreadsheet. The hotlinks at the end of this chapter have some suggestions for online questionnaire and survey websites.

In addition to following the IB's guide concerning ethical questions applied to experimentation in the lab, students interested in writing questionnaires might want

Google Forms is a no-cost solution for making an online survey. Some online services for surveys limit the number of respondents or charge a fee if you get more than 100 respondents, but this is not the case for Google Forms.

to have a look at the diploma programme guide for psychology. In it, there are clear guidelines about what the IB considers to be acceptable practice when using human subjects for an investigation.

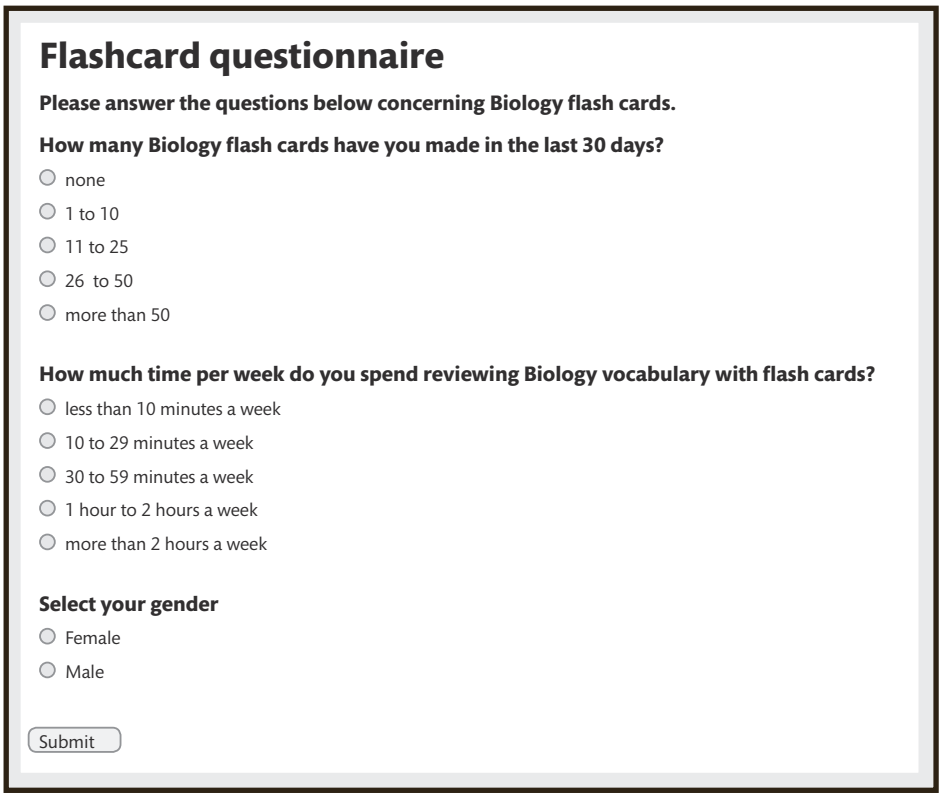


Figure 12 Example of three questions from an anonymous online questionnaire using multiple-choice questions from Google Forms. Other questions might gather data about students' results on their last biology test or whether they are taking HL or SL Biology.

Data-analysis exercises

Both for your internal assessment work and in data-based sections of exams, you will be required to interpret sets of data presented either as tables or as graphs. Being able to extract scientific information from data is a key skill in biology.

The first thing to look for on a table or a graph is a title. When titles are not available, often the text before or after the tables and graphs will reveal some key information about what they are showing. The next clues to look for in order to interpret the data correctly are labels and units in the headings of tables, or labels and units on the axes of graphs. In both cases, the labels are often the dependent and independent variables of the investigation that generated the data. Knowing these will help you reach conclusions about the investigation. The units might be familiar to you, such as grams, millilitres, or °C, but sometimes they are units you have never heard of. In such cases, do not panic, just be sure to include those unfamiliar units in your answers and in your analysis. The same goes for arbitrary units, which are sometimes used to avoid employing confusing units.

Next, look at the scales on the axes of graphs. Do they show regular intervals (10, 20, 30, 40) or is there an atypical scale, such as a logarithmic scale (1, 10, 100, 1000)? If two graphs are being compared, do they use the same scales and the same maximum and minimum values? If not, be careful with how you compare the two because they may look the same but in fact be very different.

Worked example

Analyse the graph below showing the sizes of wings of fruit flies in Europe, North America, and South America. Note that the original species of *Drosophila subobscura* lived in Europe and was introduced to the Americas in recent decades. What scientific information can be concluded from the graph? For example, can any predictions be made?

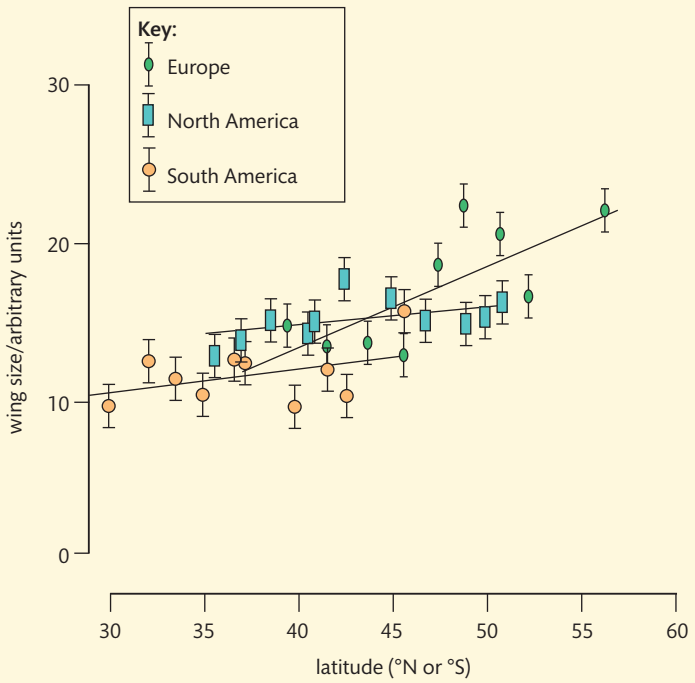


Figure 13 Graph showing fruit fly *Drosophila subobscura* wing sizes in different parts of the world.
Gilchrist et al. 2004

Solution

Reading the graph: look at the axes, showing latitude for the *x*-axis and wing size for the *y*-axis. Latitude is a measurement of how many degrees away from the equator something is: low numbers are closer to the equator, high numbers are further away from the equator. Next, look at the key: there are three different shapes and colours to analyse, depending on where the flies were observed. Associated with each group is a trend line. In addition, each data point has vertical error bars.

Analysis of the graph: all three trend lines increase as the latitude gets further from the equator, but the one from Europe has the greatest slope. If we look at the centre of each trend line, it appears that the South American population has the smallest wing size, the North American population has an intermediate wing size, and the European population has the biggest wing size. The European population has the widest range of wing sizes.

Conclusion: There is a relationship between latitude and wing size: they are positively correlated. We can predict that if the introduced populations of *D. subobscura* in the Americas were to spread to latitudes that are further away from the equator than the ones show in the graph, they should show an increase in wing size.

Throughout this book, there are examples of past paper questions, and some of them have graphs or tables of numbers that need to be interpreted and analysed. Be sure to practise analysing them because that is what you will be asked to do in exams.

Fieldwork and data logging

Biology investigations carried out in labs allow students to have a certain amount of control over the variables that they are manipulating. Fieldwork, however, does not offer such possibilities. Studying a forest, stream, grassland, or marine environment poses some unique challenges. Abiotic factors, such as temperature, air humidity, and light, can vary considerably, and could have an influence on what is being studied. For example, setting up pitfall traps is a wonderful way to collect invertebrates in a forest or grassland, but adverse weather conditions might greatly affect how active invertebrates are. Because they cannot be controlled, abiotic factors should be monitored and data should be collected to make sure that they do not have an adverse effect on the results.

A student using a hand-held data-logging device to measure the pH of a sample of water in a river.

Collecting large quantities of data can sometimes be tedious and prone to errors if done by hand. Instead of using a thermometer and writing down the temperatures, students can use temperature probes connected to data-logging devices that can automatically record temperatures at particular intervals. Such devices can be equipped with probes for:

- temperature
- light intensity
- relative humidity
- flow rate (to see how fast water is flowing)
- dissolved oxygen.

Data loggers can also have an integral global positioning system (GPS) (to record the exact location of each measurement).

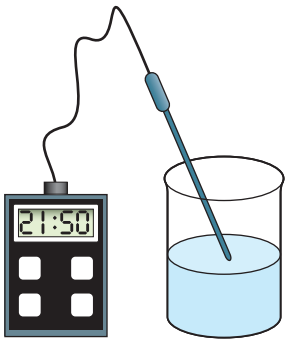


Figure 14 A hand-held data-logging device with a probe measuring the temperature of a solution in a beaker.

Once the probes are plugged in, the data-logging device can be used in various modes depending on how the data are to be collected. Here are some examples:

- real-time (useful for monitoring measurements without recording them; in this mode, the device is used as a simple meter)
- events with entry (useful for measuring a certain factor every metre, or when another event happens; data will only be recorded when you tell the device to do so)
- time-based (useful for measuring something constantly over a fixed amount of time; in this mode, parameters can be adjusted to measure every minute, hour, second or fraction of a second).

Although the data can be transferred to a computer, many of these devices allow you to graph and analyse the data directly on screen. This is especially useful when doing fieldwork without a readily available computer.

ICT skills applied to your lab reports

To produce top-quality lab reports, it is expected that students have access to the following types of programs.

- Word processing software: programs such as Microsoft Word will help you with text, tables, footnotes, and chemical as well as mathematical formulas.
- Spreadsheet software: programs such as Microsoft Excel will help you with data processing to perform calculations such as averages, standard deviation, chi-squared tests, and more.
- Graphing software: in addition to its spreadsheet functions, Microsoft Excel also has graphing capabilities, to make line graphs, bar charts, histograms, and scatter plots, to which you can add trend lines and automatically calculate correlation coefficients.

Students who do not have access to Microsoft products can find other solutions, such as cost-free software packages available from OpenOffice.org, or online applications such as the ones available from Google. See the hotlinks section at the end of the chapter for more information.

The sections below list the functions and capabilities of the various software programs that students should consider learning about and using in their lab reports. To find out how to use them, look through the menus of whichever program you are using. The way the lists are set out below, the first word suggests which menu or tab to start to looking in, although many programs have icons with some of the more frequently used functions. Examples have been provided for some functions.

Word processing

Students should know how to do the following things with a word processor.

- Format: changing text formatting, such as putting species names in *italics*.
- Format: turning numbers into subscripts (e.g. H_2O) and superscripts (e.g. cm^3).
- Table: setting up tables, merging cells, aligning text horizontally/vertically within cells, rotating text 90°.
- Table/ruler/tabs: aligning decimal points within columns of a table.
- Table: adding borders around the cells so that they show up clearly.
- Insert: adding bulleted lists and numbered lists.

If you cannot find a feature in a program you are using, do not hesitate to go online. Do a search for 'how do I ...?' and type in the function you are looking for, and finish the search with 'in ...' and type in the name of the software and the version. If it is important that the solution is specifically for Mac, say so in your search otherwise there is a good chance the solutions you find will be for non-Mac users. For example, 'How do I insert footnotes in Microsoft Word for Mac?'

- Insert: adding a photo and resizing it to fit, and including a legend with the photo. Students should know how to adjust the quality or resolution of images to avoid the problem of the file size of their documents being too big. This can be a particular issue when submitting a document electronically.
- Edit: pasting a graph copied from a graphing program; if the lab report is going to be submitted electronically, is best to paste the graph as an image rather than as a linked object.
- Insert: using shapes such as arrows or boxes to annotate an image.
- Insert: adding notes such as footnotes at the bottom of a page, or endnotes at the end of the document.
- Insert: adding formulas using formula editors to produce well-presented equations to show how you processed your data. Note that some versions of word processors do not have the formula editors pre-installed so they need to be added manually.
- Insert: adding symbols such as ±, Δ, λ, or ≤ where necessary.
- Insert: using page breaks to avoid having a section start at the bottom of a page or to avoid having a table split over two pages.
- Tools: selecting the text and setting the proofing language for the language you are using.
- Edit: using paste special for pasting text or numbers without the formatting.

Spreadsheets

Students should know how to do the following things with a spreadsheet program.

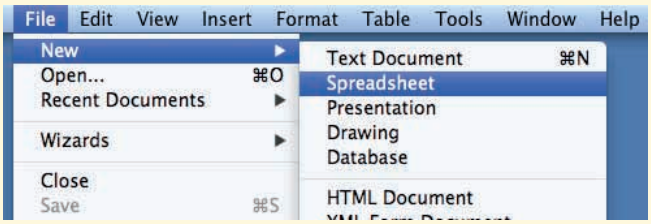
- Understand the system of identifying cells as A1, B2, C3, etc. (see screenshot 2 on the next page).
- Format: changing the format of the cell to match the type of data, such as number, date, percentage, text, time, scientific notation, etc.
- Format: changing the number of decimal places after the decimal point to correspond to the desired degree of precision.
- Insert: using math operations by inserting an equals sign '=' followed by a formula using 'A1 + A2' to add, or 'B3/B2' to divide, or '(A1 + A2 + A3)*B1' to combine more than one operation in the same formula.
- Insert: inserting predefined formulas such as sum, average, maximum or minimum, standard deviation, chi-squared, etc. Example for Excel: typing '=max(A1:A100)' in cell A101 finds the maximum value between A1 and A100. Replacing the term 'max' with 'min' in the formula finds the minimum value. Note: if your software is installed in a language other than English, the commands may be different. For example, 'sum' is 'somme' in French versions of spreadsheet software.
- For a repeating operation, copying the formula down a column or across a row rather than re-typing it separately each time.
- Converting a relative reference into an absolute reference by adding $, for example B2 does not behave the same way as $B2 or B$2 or B2 when it is copied and pasted to another place on the sheet.
- With international settings, the decimal point can sometimes be a full point (.) and sometimes be a comma (,) so if the decimals in your data do not seem to be recognized by the program, it is possible that you need to switch from one to the other. Instead of doing this manually, use the find and replace feature in the edit menu.

Worked example

Use a spreadsheet program to calculate the mean, mode, and median of the data mentioned earlier in the chapter.

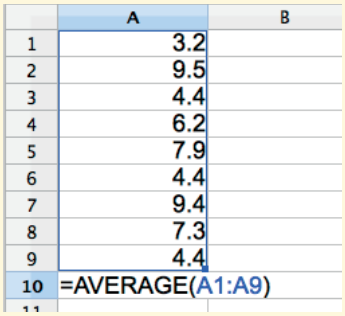
Solution

Be sure a spreadsheet program is installed on the computer or tablet you are using, such as Microsoft Excel or the spreadsheet programs available in software packages such as NeoOffice, LibreOffice, and Apache OpenOffice. The screenshots in this chapter are from OpenOffice Calc, which is available at no cost online.



Screenshot 1. Creating a new spreadsheet in OpenOffice.

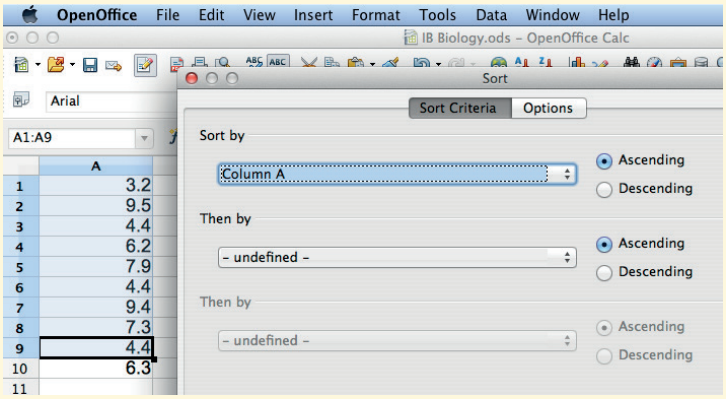
By typing in the values into the cells A1–A9 of your spreadsheet, you can then enter a formula to calculate the mean in cell A10. See screenshot 2.



Screenshot 2. Calculating the mean (average) by using the '=average' function and either selecting cells A1–A9 by manually selecting them or typing A1:A9 in the parentheses. When you hit the ENTER key, it should calculate 6.3.

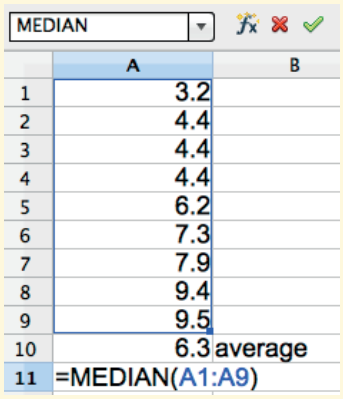
If you get an error message, be sure your program is not expecting a comma (,) instead of a full point (.) for the decimal point. In certain international versions of spreadsheet programs, the default is for a comma.

To find the median in the spreadsheet program, first select the nine values, then go to the Data menu and select Sort.



Screenshot 3. Using the sort feature to put numbers in order. Notice how cell A10 is purposely left out of the selection, as it is not part of the data (it is the calculated mean).

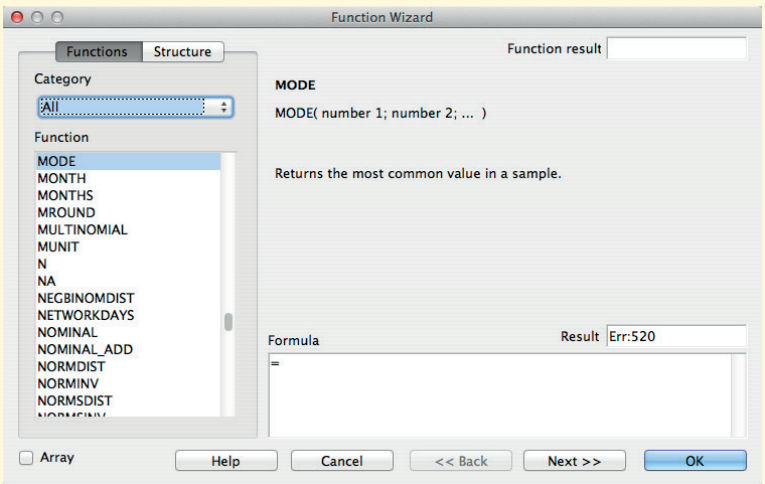
You could find the median manually by counting the data points and finding the one in the middle, or you could let the computer calculate it: type in the formula '=median(' and select the data between A1 and A9.



Screenshot 4. Using other formulas for the data.

Note that the median function works even if you do not sort the data as we did in screenshot 3.

Screenshot 5 shows finding the mode.



Screenshot 5. By selecting 'Function' from the 'Insert' menu, you can choose from a long list of possible calculations to perform. Here, 'MODE' is shown, but there are many others, such as MAX, MIN, etc. As for the other functions, use the range of data A1 to A9 in the parentheses. Microsoft Excel and other spreadsheet programs have similar lists of functions to insert.

Graphing

Students should know how to do the following things with graphing software.

- Entering the data in proper columns and rows so that the computer recognizes the data with its headings.
- Defining which data will be graphed by carefully selecting the correct rows and lines (it is important that there are no blank rows or blank columns in the selected data).
- Insert: selecting a type of graph that will lead to useful analysis.
- Insert: once the data points on a scatter plot are selected, a trend line can be inserted.
- Options: once the data points on the graph are selected and a trend line added, graphing programs often suggest options such as inserting the formula for the trend line or calculating the r^2-value.
- Options: once data points on the graph are selected, error bars can be added.

- Format: adding a title to the graph as well as labels to the axes. Many graphing programs suggest a legend, but legends are only necessary if two or more colours are used.
- Options: Once the numbers on the *x*- or *y*-axis are selected, it is often possible to alter the maximum and minimum values (useful for zooming in on a part of the graph that is interesting) or changing the scale (sometimes it is clearer to show every fifth value or every tenth value rather than every number on the scale).
- Options: for graphs showing two values for *y* measured in two different units, it is sometimes necessary to add a second *y*-axis using a different scale to avoid the problem of one variable's graph being squashed and unreadable.

As with all skills, it will probably take some time to learn the software the first time you use it. But with practice and perseverance, you should become proficient. Learning these skills will undoubtedly help you in your future studies after IB, and many will help you later in your career.

Worked example

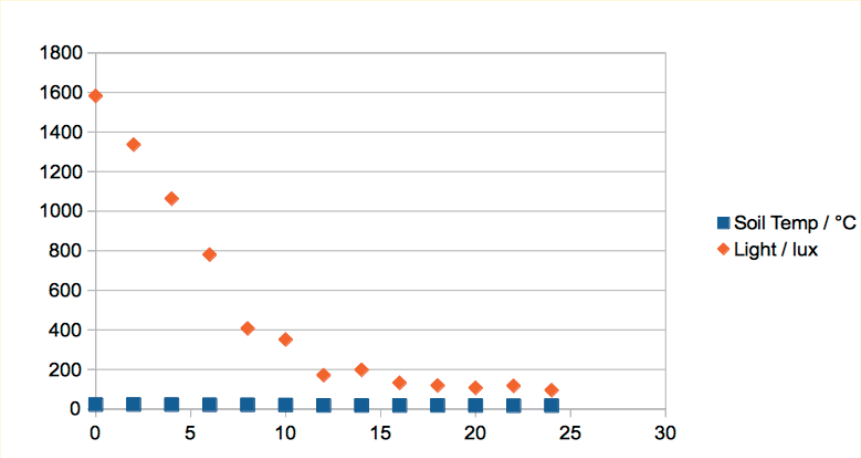
Make a scatter plot graph using the following data points of abiotic factors measured by students doing fieldwork between an open grassy area (towards the 0 m side) and a woodland (towards the 24 m side).

	A	B	C
1	**Distance / m**	**Soil Temp / °C**	**Light / lux**
2	0	22.5	1582
3	2	23.0	1336
4	4	22.4	1063
5	6	21.4	780
6	8	21.0	407
7	10	20.2	351
8	12	18.4	171
9	14	18.2	198
10	16	17.8	132
11	18	18.3	119
12	20	17.8	107
13	22	16.7	117
14	24	17.4	95

Screenshot 6. The raw data.

Solution

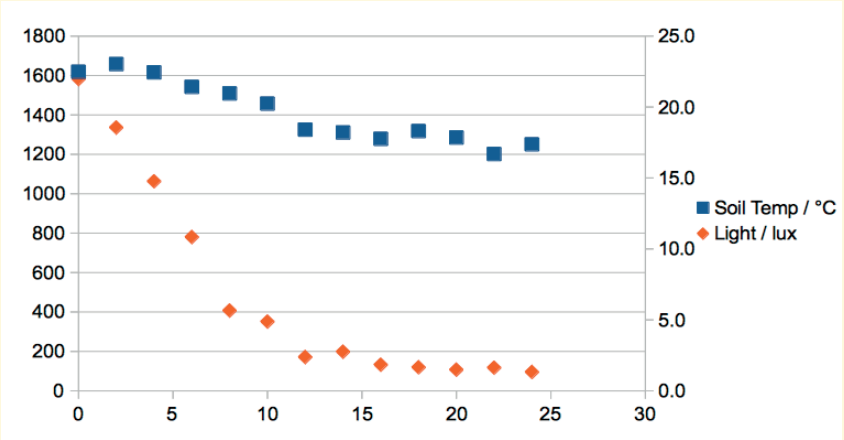
First, enter the data as shown in columns A, B, and C. Second, select the cells A1–C14: it is recommended that you include the labels of the data in row 1 when graphing data. Note: too many students look at data like this and decide to do two separate graphs, one for soil temperature and another for light levels. However, it saves space and allows a better comparison if both variables are graphed together. Third, indicate to your spreadsheet program that you want to insert a graph. Fourth, choose the graph type, in this case a scatter plot, in order to plot *y* against *x*.



Screenshot 7. A graph that is not very useful.

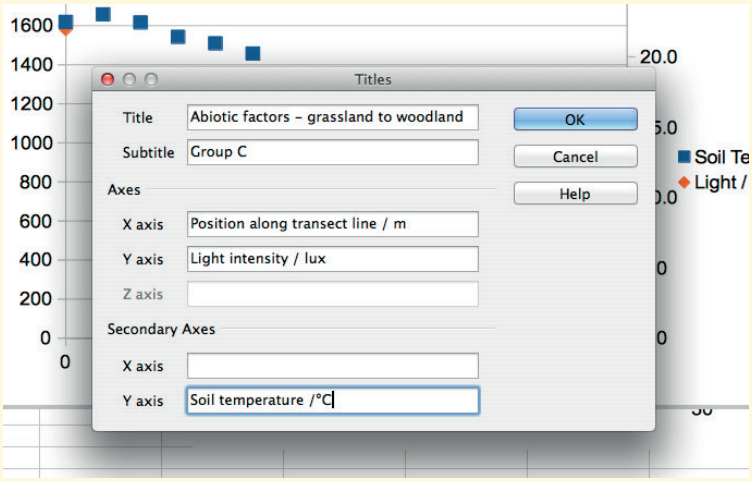
What you have at this point (screenshot 7) is far from satisfactory, and you need to work through quite a few options in order to obtain a graph that will allow you to analyse the trends.

By selecting the blue data points (the soil temperature), it is possible to right-click on them and ask OpenOffice to change the format of the data series. Select 'secondary Y axis' for the blue data points. This will create a second *y*-axis on the right.



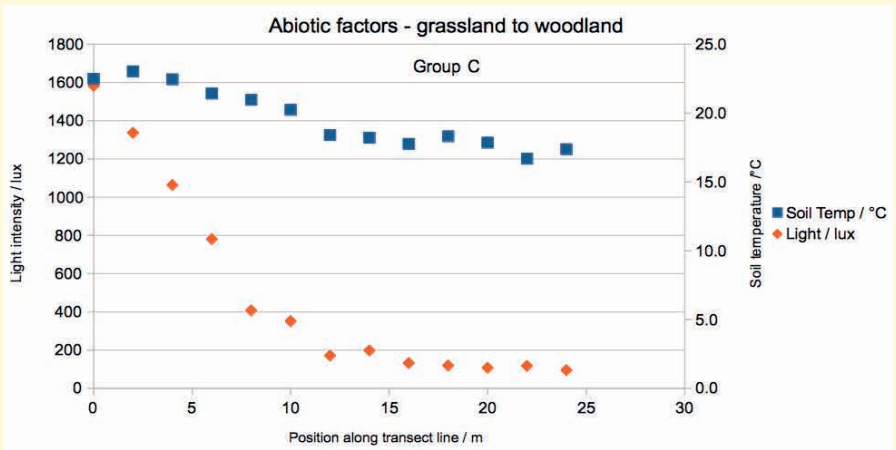
Screenshot 8. Creating a secondary *y*-axis.

This solves the problem of not being able to see any changes in the soil temperature in °C because it is on the same scale as the light readings in lux. But the graph is still not finished. To add a title to the graph and labels on the axes, make sure the graph is selected (by double-clicking on it) and choose 'Titles' from the options.



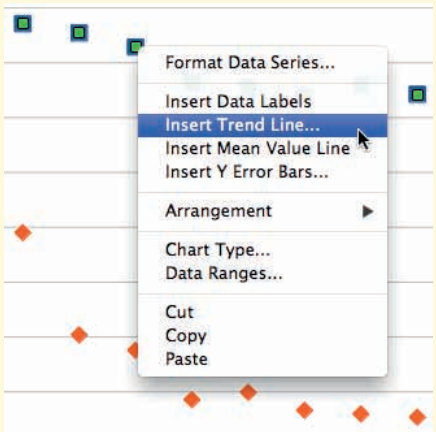
Screenshot 9. Adding a title and labels for the axes.

Now the graph should look like screenshot 10.



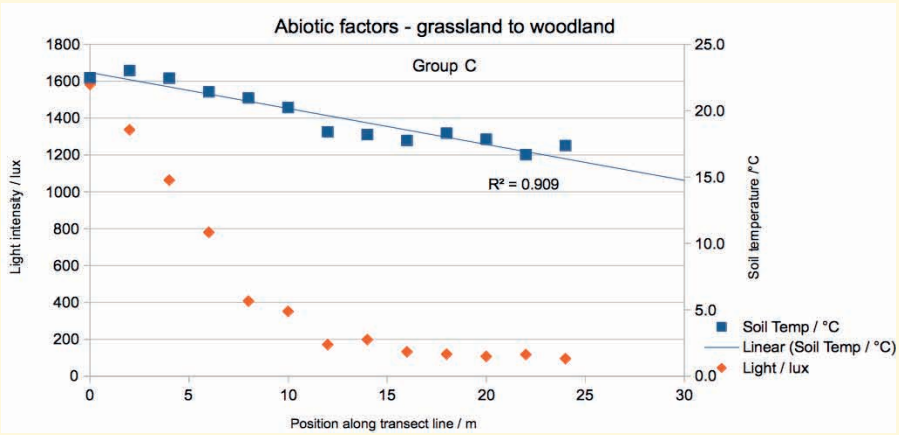
Screenshot 10. Labels added.

To finish data processing, a trend line could be added to one or both sets of data points. Here is an example of a trend line being added to the soil temperature by right-clicking on the selected data points.



Screenshot 11. Using a drop-down menu to insert a trend line.

The r^2-value can been added by right-clicking on the selected trend line and choosing the format options. Do not be surprised if the graphing software writes r^2 as R^2.



Screenshot 12. The graph is now ready to interpret.

There is a trend in the temperature that goes from warmer temperatures in the open grassland near the 0 m mark, and cooler temperatures in the woodland towards the 24 m mark. The light levels seem to be a good indication of where the tree line starts to block out the sunlight after 10 m. This graph might inspire you to see the correlation from 0 to 10 m and compare it to 11 to 20 m.

To learn more about maths and ICT skills, go to the hotlinks site, search for the title or ISBN and click on Mathematics, and information and communication technology skills.

Just because you spend time making a graph look great does not mean it is worth including in a lab report. Sometimes it will inspire you to look for other patterns. In this example, further processing could be done by plotting soil temperature against light levels to see whether they are correlated. It is advisable that you not wait until the night before an assignment on data processing is due before learning how to graph. Remember when you were a child and you first learned how to tie your shoelaces? Remember how long it took the first time? This is true for many skills and preparing graphs for data processing is no exception. The first few times you make a graph will take a long time. Once you are an expert, it will take much less time.

Here is a closing thought about your smartphone: have you ever considered using it as a measuring device for lab work? Smartphones have microphones to measure sound levels, accelerometers that can be used for detecting vibrations or measuring angles as a spirit level, a camera that can be used as a lux meter or to make slow motion videos, a GPS that can measure your position and many other things. Check out app stores online for the ones available for your phone. Ideas include a click counter, decibel meter, timer, tape measure, 'radar' speed gun, and many more. When doing microscope work in the lab, certain smartphone cameras work quite well for taking photos of what you are observing. Different teachers have different philosophies about the use of smartphones in the lab or during fieldwork but it is a shame not to take advantage of this powerful computer in your pocket. Whether or not you become a scientist later in life, the maths and ICT skills you learn in the IB will help you throughout your life.